The contribution of perceptual factors and training on varying audiovisual integration capacity

Article (Accepted Version)

Running head: Varying audio-visual integration capacity

The contribution of perceptual factors and training on varying audio-visual integration capacity

Jonathan M. P. Wilbiks [1, 2]

Benjamin J. Dyson [1, 3]

[1] Ryerson University, Toronto, ON, Canada

[2] Mount Allison University, Sackville, NB, Canada

[3] University of Sussex, Falmer, UK

Address : Department of Psychology

Mount Allison University

62 York Street

Sackville, New Brunswick

Canada

E4L 1E2

Telephone : +1 (506) 364-2269 ext 2458

E-mail : jwilbiks@mta.ca

Abstract

The suggestion that the capacity of audio-visual integration has an upper limit of 1 was challenged in four experiments using perceptual factors and training to enhance the binding of auditory and visual information. Participants were required to note a number of specific visual dot locations that changed in polarity when a critical auditory stimulus was presented, under relatively fast (200 ms SOA) and slow (700 ms SOA) rates of presentation. In Experiment 1, transient cross-modal congruency between the brightness of polarity change and pitch of the auditory tone was manipulated. In Experiment 2, sustained chunking was enabled on certain trials by connecting varying dot locations with vertices. In Experiment 3, training was employed to determine if capacity would increase through repeated experience with an intermediate presentation rate (450 ms). Estimates of audio-visual integration capacity (*K*) were larger than 1 during cross-modal congruency at slow presentation rates (Experiment 1), during perceptual chunking at slow and fast presentation rates (Experiment 2), and, during an intermediate presentation rate post-training (Experiment 3). Finally, Experiment 4 showed a linear increase in *K* using SOAs ranging from 100 to 600 ms, suggestive of quantitative rather than qualitative changes in the mechanisms in audio-visual integration as a function of presentation rate. The data compromise the suggestion that the capacity of audio-visual integration is limited to 1 and suggest that the ability to bind sounds to sights is contingent on individual and environmental factors.

KEYWORDS: audio-visual integration; capacity; multimodal perception; training;

Public Significance Statements

This study strongly suggests that the capacity of audiovisual integration is a flexible structure, and that this capacity can increase as a function of specific stimulus factors.

This study indicates that training can be effective in increasing the capacity of audiovisual integration (although this training does not transfer to differing speeds of presentation).

This study shows that, audiovisual integration capacity modulates quantitatively, rather than qualitatively, as a function of stimulus onset asynchrony.

Introduction

While navigating our everyday lives, we are constantly stimulated by various sensory inputs in several different modalities, some of which are ultimately perceived as unified multi-modal events. Welch and Warren (1980) present multisensory integration as an implicit decision-making process, wherein an individual must decide whether two sensory inputs they receive are caused by the same event or multiple different events. Whether integration occurs or not is based on a number of factors (see Koelewijn, Bronkhorst, & Theeuwes, 2010 for a more comprehensive review), including timing and cross-modal congruency. In terms of temporal factors, there is a range of timing within which an auditory and a visual stimulus is more likely to be bound, referred to as the *temporal window of integration* (TWI). At its most basic, the TWI extends from around 30 ms auditory lead to around 170 ms visual lead in sensation (Van Wassenhove, Grant, & Poeppel, 2007). This asymmetry is likely a consequence of the faster transduction of light relative to sound in the atmosphere. While specific estimates of this temporal window vary (e.g., Zampini, Shore & Spence, 2003; Spence & Squire, 2003; van Wassenhove, Grant, & Poeppel, 2007; Soto-Faraco & Alsius, 2009), most research finds that audio-visual binding between two stimuli is optimized when the visual stimulus occurs around 85-100 ms ahead of an auditory stimulus. Moreover, the window of integration has been shown to be flexible both between (Fujisaki, Shimojo, Kashino, & Nishida, 2004; Heron, Whitaker, McGraw, & Horoshenkov, 2007) and within individuals (Stone, Hunkin, Porrill, Wood, Keeler, Beanland, Port, & Porter, 2001).

Congruency factors have also been shown to influence likelihood of binding, with related auditory and visual stimuli shown to be more likely to be bound (Spence, 2011). Spence puts forth three general types of cross-modal correspondences: structural, statistical, and semantically mediated correspondences (although see Walker, 2012, for an alternative view). Structural correspondences are those which occur due to "intrinsic attributes of the perceptual system's organization" (Spence, 2011, p. 988). That is to say, if certain unimodal stimulus traits are processed in proximal areas in the brain, there is likely to be a correspondence between those traits (Ramachandran & Hubbard, 2001). Walsh's (2003) ATOM (A Theory of Magnitude) theory, proposes that there is a common coding, and hence structural congruency, between auditory loudness and visual brightness. Statistical correspondences are based on regularities in the environment, and our subsequent exposure to these regularities leading to an increased correspondence between two stimuli. For example, since the resonance properties of objects require that a small object generate a high-pitched sound, there is a cross-modal correspondence between high pitch and small size (and low pitch with large size; and see also Marks, 1987; Evans & Treisman, 2010). Finally, semantically mediated correspondences relate to the use of common language to describe different sensory inputs. For example, shared use of "high" and "low" verbal labels contribute to cross-modal correspondence between auditory pitch and visual height (Rusconi, Kwan, Giordano, Umilta, & Butterworth, 2006; Leboe & Mondor, 2007).

As well as studying the way in which factors such as temporal coincidence and stimulus congruency work with one another to resolve binding across modalities (e.g.,

Wilbiks & Dyson, 2013a, b), attention has more recently turned to the *quantity* of information that can be integrated across vision and audition. Capacity limits are central information processing constructs in both uni-modal and multi-modal literatures. Well-known examples include the capacity of visual short term memory (VSTM; Cowan, 2001; Todd & Marois, 2004; Vogel, McCollough, & Machizawa, 2005) that is currently estimated around 3 or 4 visual items, and, the limit of semantic working memory thought to revolve around the 'magic number' 7 +/- 2 (Miller, 1956). More recently, attention has turned to estimating the upper bound of audio-visual integration capacity. From an ecological point-of-view, it is reasonable to believe there should be a limit (Van der Burg, Awh & Olivers, 2013), and since there is generally a single visual object that shares causality with a single auditory event (e.g., one panda generates one sneeze), that limit should be 1. To provide empirical support for this position, Van der Burg et al. (2013) presented participants with 16 or 24 dots, arranged along an imaginary circle, of which up to 8 changed polarity from black to white (or vice versa) repeatedly at an SOA of 150 or 200 ms. Critically, on one of the presentation frames an auditory signal was presented. Participants were provided with a probe location and were required to indicate whether that specific location had changed polarity on the frame where the auditory signal was heard. By submitting their accuracy data to curve fitting they produced estimates of audio-visual integration capacity, and found that across all of their conditions the capacity of audio-visual integration was never greater than 1 item. Participants were only able to reliably bind no more than one visual location with the auditory tone. Olivers, Awh, and Van der Burg (2016) provided further support for this hypothesis by showing that in

addition to participants only being able to track one changing dot, they were also only able to reliably report the orientation of a line overlaid on a single dot.

In a recently published paper, Olivers et al. (2016) set out their *single source hypothesis*, which holds that in an audio-visual binding scenario, binding is limited to a maximum of one auditory-visual pairing. As such, the two stimuli contributing to that pair are encoded with high precision, regardless of the number of other stimuli that are present. Increasing set size leads to a decrease in likelihood of correct binding, but does not decrease the precision of detail that one can report when one is correct. Consequently, this hypothesis argues against a distribution of attention across multiple stimuli in favour of a mechanism where a single, bound stimulus is processed both in general and in detail. The single source hypothesis is also reminiscent of research in the uni-modal literature, specifically with the feature binding literature. Research in this field suggests that binding between stimuli tends to occur on at a one-to-one ratio, regardless of whether the bound pair is between two stimuli or between one stimulus and one response (Frings, Rothermund, & Wentura, 2007). However, it is also useful to consider additional research in the feature binding literature, which holds that three or more stimuli can be bound with one another through concurrent pairings (e.g. A bound to B while B is also bound to C; Hommel, 1998; Hommel & Colzato, 2004; Hommel, 2004).

Statements about the strict upper limits of audio-visual integration capacity run counter to additional observations both within the uni-modal and multi-modal literatures, in which capacity varies both as a function of environmental and individual factors. In addition to the observation of multiple bindings in the uni-modal domain (Hommel,

2004), there is clear variation in the range of values reported for VSTM capacity (e.g., 1.5 – 6; Vogel & Machizawa, 2004). Similarly, while the group average capacity for audio-visual integration in Van der Burg et al.'s (2013) studies did not exceed one item, they found individuals for whom capacity was higher than one (e.g., range 0.70 to 1.56; Experiment 1c). Second, the capacity of audio-visual integration appears subject to some of the same factors that also improve VSTM capacity. For example, additional experiments in Van der Burg et al. (2013) showed that the reduction of visual perceptual load (e.g., Lavie, 2005) from 24 to 16 items led to a non-significant increase in performance, whereas slowing down the rate of visual presentation from 150 to 200 ms (e.g., Holcombe and Chen, 2013) led to a significant increase in the number of visual locations that could be successfully tracked.

In a similar series of experiments, Wilbiks & Dyson (2016) adopted a version of the Van der Burg et al. (2013) paradigm in which the perceptual load of the task was further reduced to 8 elements, and a more extreme manipulation of the speed of visual presentation was deployed (200 versus 700 ms). As acknowledged by Van der Burg et al. (2013, p. 348) very fast SOAs (e.g. 200 ms) are more likely to lead to potential mis-bindings between vision and vision, assumedly due to the auditory stimulus falling within the TWI of multiple visual stimuli. Therefore, increasing SOAs implies less susceptibility to mis-bindings, leading to a potential increase in the capacity of AV integration. Furthermore, the temporal predictability of the critical audio-visual binding event (e.g., Wasserman, Chatlosh, & Neunaber, 1983) and the level of proactive interference incurred by previous visual frames (e.g. Luck & Vogel, 1997) were also

manipulated to understand the conditions under which audio-visual integration capacity might be malleable. A critical finding in these experiments was that audio-visual integration capacity could exceed 1, but this appeared to be limited to slower rates of presentation (i.e., 700 ms). We also included a visual-only control condition (Wilbiks & Dyson, 2016; Experiment 5), which revealed no facilitatory effect of visual-visual binding at any SOA. This suggests that audio-visual integration was occurring at both 200 and 700 ms in the series of experiments, since the inclusion of a temporally aligned visual cue was not sufficient to trigger integration. To further assist in answering whether the capacity of audio-visual integration can exceed 1, we pursued both transient (cross-modal correspondence; Experiment 1) and sustained (chunking; Experiment 2) perceptual manipulations hypothesized to aid multi-modal capacity. In Experiments 3 and 4, we addressed whether these increases in capacity represented qualitative or quantitative changes in mechanisms of multi-modal processing. This was achieved by examining the effects of training (Experiment 3) and by evaluating performance across a wider range of SOA (100 to 600 ms; Experiment 4).

Experiment 1

When considering current iterations of the audio-visual integration task (Van der Burg et al., 2013; Wilbiks & Dyson, 2016), the way to maximize performance is to successfully bind as many visual candidates as possible to the auditory stimulus, in hope that one of those candidates is the one that is eventually probed. Put another way, there is no unique discriminating information between changing dot locations at the time of the critical trial: all visual stimuli that change do so simultaneously, in locations that are all

equidistant from fixation, and, deploy polarity changes of equal salience (black to white, or, white to black). The inspiration for Experiment 1 is the large literature showing that cross-modal congruency can promote multimodal binding (see Spence, 2011, and Walker, 2012). These congruency relationships can have their root in certain strcutural commonalities such as size and pitch (Gallace and Spence, 2006), or in more abstract factors based such as height and pitch (Parise and Spence, 2009), which are determined either by second-level, statistical correspondences (Spence & Deroy, 2012) or by semantic labels (Walker, 2012).

In terms of the impact of cross-modal correspondence, congruent relationships between visual and auditory information can both increase perceptual sensitivity and attentional capture. For example, Marks, Ben-Artzi, and Lakatos (2003) found that congruent cross-modal stimuli increased perceptual sensitivity on both auditory and visual stimulus discrimination tasks. While perceptual sensitivity to a stimulus in one modality was increased in the presence of a stimulus in the other modality, this effect was not symmetrical: there was a stronger effect found for an auditory stimulus accompanying visual perception, relative to a visual stimulus accompanying auditory perception. Therefore, the current paradigm is well positioned to take advantage of the strong effect of an auditory stimulus introduced during continuous visual perception. Also importantly for the promotion of audio-visual integration capacity, cross-modally congruent stimuli lead to increases in attentional capture across modalities (Shams & Kim, 2010), with an auditory stimulus increasing attention to a congruently paired visual stimulus (see also Fiebelkorn, Foxe, and Molholm, 2010). As a result of these

observations, we also expect cross-modally congruent stimuli to increase the capacity of audio-visual integration.

To this end, Experiment 1 employed a factor of pitch-brightness congruency by manipulating the pitch of the tone deployed at the critical trial. During valid trials in Van der Burg et al. (2013) and Wilbiks & Dyson (2016), the to-be-probed location could switch between two states: white to black, or, black to white. Marks (1987) found that light coloured (e.g. white) visual stimuli were congruent with high-pitched tones, and that dark coloured (e.g. black) visual stimuli were congruent with low-pitched tones, while Parise and Spence (2009) showed that cross-modally congruent stimuli using appropriate combinations of brightness and pitch also increased the temporal window of integration between two stimuli. We expected that the presentation of a low tone during the critical trial would promote binding to locations that changed from white to black, whereas the presentation of a high tone during the critical trial would promote binding to locations that changed from black to white. Therefore, the capacity of audio-visual integration should be higher during congruent relative to incongruent trials.

Method

Participants. All experimental and recruitment practices were approved by the Research Ethics Board at Ryerson University. 24 participants were recruited from an undergraduate research participant pool, and compensated with partial class credit. After Wilbiks & Dyson (2016), we calculated a 95% confidence interval (CI) around 50% and removed 4 participants who performed within that CI, on average and across all conditions. The final sample consisted of 20 participants with an average age of 20.8,

with 16 females and 18 right handed individuals. This sample size was deemed appropriate as the observed effect size of the highest level interaction from Wilbiks & Dyson (2016) was $\eta_p^2 = .177$. Assuming a similar effect size, with $\alpha = .05$ and Power $(1 - \beta) = .80$, it was determined that the two way ANOVA conducted in Experiment 1 required a minimum sample size of 20 participants (G*Power 3; Faul, Erdfelder, Lang, & Buchner, 2007).

Stimuli. Visual stimuli were presented on a Viewsonic VE175 monitor at a viewing distance of approximately 57 cm. Stimulus generation and delivery was controlled by Presentation software (version 16.5, build 09.17.13), Visual stimuli consisted of dots 1.5° in diameter displayed in either black (0, 0, 0) or white (255, 255, 255) against a mid-grey background (128, 128, 128). Eight dots at a time were presented along an implied circle, which had a diameter of 13°, the center of which was marked by a 0.15° fixation dot. A single, smaller probe dot was overlaid on a target dot at the end of each trial, and was red (255, 0, 0) with a diameter of approximately 1°. Auditory stimuli were created using SoundEdit 16 (MacroMedia) and consisted of a 60 ms tone with 5 ms linear on-set and off-set ramps, either low (300 Hz) or high (4500 Hz) in pitch (after Parise & Spence, 2009). Sounds were presented binaurally via Sennheiser HD 202 headphones at an intensity of approximately 74 dB(C).

Design and Procedure. Experiment 1 was based upon Wilbiks & Dyson (2016; Experiment 4). 16 individual conditions were created, by orthogonally varying the SOA of visual stimuli (200 or 700 ms), the number of visual stimuli changing on each alternation (1, 2, 3, or 4), the cross-modal congruency of the to-be-probed dot and the

tone (congruent or incongruent). These 16 conditions were each presented 4 times (2 valid probes, 2 invalid probes) to create an experimental block with 64 trials. Each participant completed one practice block of 16 trials, and 6 experimental blocks consisting of 64 trials each, for a total of 384 experimental trials. Figure 1a provides a schematic of Experiment 1. For valid trials where the to-be-probed dot changed polarity at the critical frame, a trial was deemed to be cross-modally congruent either when the target dot changed from black to white in synchrony with a high-pitched tone, or, changed from white to black in synchrony with a low-pitched tone. For invalid trials where the to-be-probed dot did not change polarity at the critical frame, there was an equal chance that the probed location colour was congruent or incongruent with the tone.

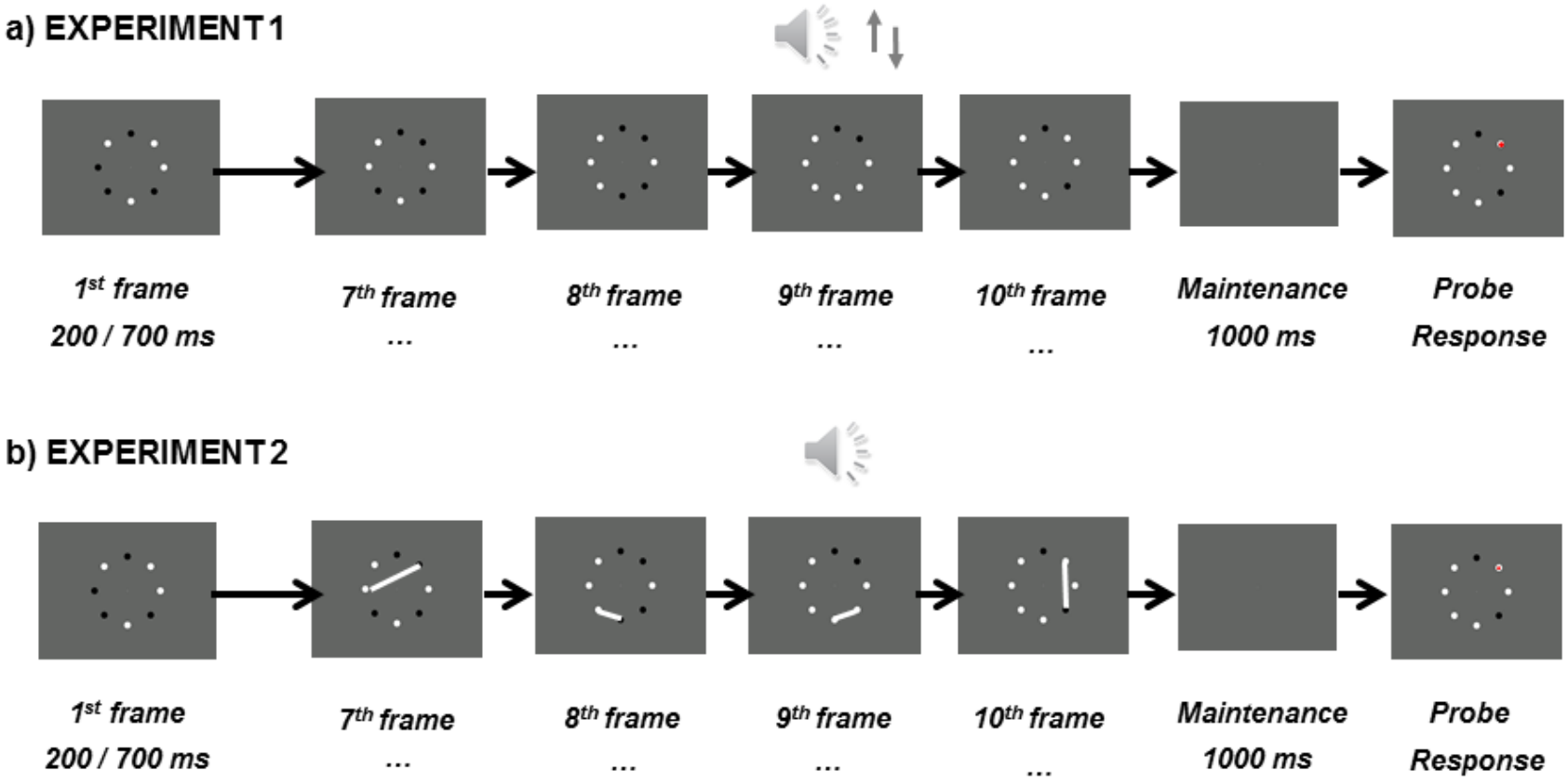

Figure 1. Trial schematics for Experiments 1 – 2.

Each trial began with the fixation point displayed in the center of the screen for 500 ms. Ten sets of eight (black or white) dots were generated for each frame and presented for either 200 or 700 ms (dependent on SOA), for a total of ten presentations. On each of these presentations, a subset of the dots (as determined by the experimental design, as above) changed polarity from black to white or from white to black. On the penultimate (9^{th}) presentation, the onset of the dots was accompanied by an auditory tone that was either congruent or incongruent with the critical location that would subsequently be probed. Following a 1000 ms retention interval, the 9^{th} array of dots was displayed again, along with an overlay of a red probe dot on one of the dots. The probe was presented visually only as it was not meant to be a memory cue, but rather a location cue for responding. Participants were asked to respond to whether the dot at the probe location had changed or not on the critical frame using a keypad. No feedback was provided, and the subsequent trial began immediately after a response was entered. Trial order was randomized in practice and in experimental trials.

Results

Estimates of audio-visual capacity (*K*) were derived in the same manner as employed by Van der Burg et al. (2013) and Wilbiks & Dyson (2016), following a variant of Cowan's *K* (2001). This model holds that if the number of locations changing is less than the capacity under given conditions, proportion correct should be maximal (if $n \leq K$, then $p = 1$). If, however, the number of locations changing is greater than capacity, expected proportion correct can be calculated as a function of both capacity and chance

(if $n > K$, then $p = K/2n + .5$). Successful model fit was confirmed by the low RMSEs observed (range 0.0001 – 0.1198). Capacity measures (K) were entered into a 2 x 2 within-participants ANOVA, with the factors congruency (incongruent, congruent) and SOA (200 ms, 700 ms). The results are displayed in the upper panel of Figure 2. While we see diversity in individual response patterns, data was analyzed on a group level with individual participant data shown in figures in the interest of transparency. This analysis revealed a main effect of congruency: $F(1,19) = 16.41$, MSE = 0.092, $p < .001$, $\eta_p^2 = .463$, with capacity for crossmodally congruent pairings yielding a significantly higher capacity than incongruent pairings. There was also a main effect of SOA: $F(1,19) = 52.95$, MSE = 0.280, $p < .001$, $\eta_p^2 = .736$, with higher capacity for slow relative to fast SOA. These two main effects were subsumed by a significant congruency x SOA interaction: $F(1,19) = 13.61$, MSE = 0.059, $p = .002$, $\eta_p^2 = .417$. Tukey's HSD ($p < .05$) confirmed that congruent audio-visual relationships significantly increased capacity relative to incongruent presentation at the 700 ms SOA (1.30 versus 1.78) but not at the 200 ms SOA (0.64 versus 0.72). To assess the conditions under which AV capacity exceeded 1, estimates of K for the four conditions were submitted to single sample t-tests against 1. Capacity remained significantly less than 1 for both incongruent ($t[19] = -4.01$, $p < .001$) and congruent ($t[19] = -3.19$, $p = .005$) presentations at 200 ms SOA. At 700 ms SOA, capacity was no different from 1 during incongruent presentation ($t[19] = 2.02$, $p = .058$) but exceeded 1 during congruent presentation ($t[19] = 4.93$, $p < .001$).

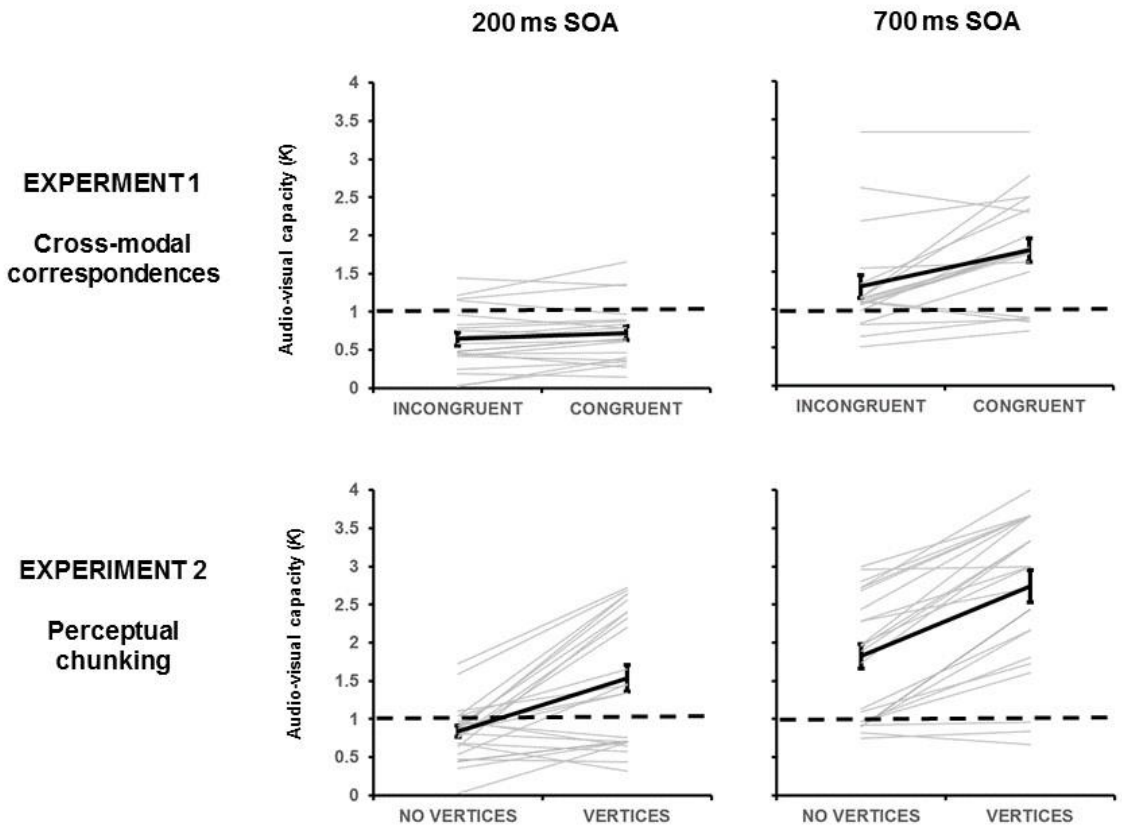

Figure 2. Audio-visual integration capacity (K) as a function of SOA (200 or 700 ms) and cross-modal (in)congruency between colour switch and tone pitch at the critical trial (Experiment 1), and presence of perceptual chunking mechanisms (Experiment 2). Grey lines represent individual participant data, whereas black lines and error bars represent means and standard errors.

In order to further delineate the conditions under which audiovisual integration is maximized, we examined the proportion of correct responses for each SOA, number of locations changing, and level of cross-modal congruency (means and standard errors displayed in Figure 3). The data were submitted to a 2 (SOA: 200 ms, 700 ms) x 2 (congruency: incongruent, congruent) x 4 (number of locations: 1, 2, 3, 4) repeated measures ANOVA. This analysis revealed expected main effects of SOA ($F(1,19) = 90.09$, $MSE = 0.019$, $p < .001$, $\eta_p^2 = .826$) and congruency ($F(1,19) = 17.74$, $MSE = 0.007$, $p < .001$, $\eta_p^2 = .483$), in addition to an unsurprising effect of number of locations ($F(3,57) = 439.66$, $MSE = 0.002$, $p < .001$, $\eta_p^2 = .959$). There were also significant interactions between SOA and congruency ($F(1,19) = 10.21$, $MSE = .004$, $p = .005$, $\eta_p^2 = .349$), SOA and number ($F(3, 57) = 4.04$, $MSE = .005$, $p = .011$, $\eta_p^2 = .175$), and congruency and number ($F(3, 57) = 7.20$, $MSE = .001$, $p < .001$, $\eta_p^2 = .275$), all of which were subsumed in a three-way interaction between SOA, congruency, and number of locations changing ($F(3, 57) = 12.31$, $MSE = 0.001$, $p < .001$, $\eta_p^2 = .393$). This three-way interaction was probed by means of a Tukey's HSD ($p < .05$) post-hoc test. Congruency was shown to have a significant facilitatory effect on response accuracy for 700 ms SOA, and when 2, 3, or 4 locations were changing, similar to the effect shown in capacity estimates in Wilbiks & Dyson (2016).

Discussion

Experiment 1 showed that the capacity of audio-visual integration can exceed 1 at a group level when slow rather than fast rates of visual presentation were deployed, and, when there was a pitch-brightness correspondence (Marks, 1987; Parise & Spence, 2009)

between the auditory signal and visual location at the critical frame. The significant interaction revealed a potentially critical constraint in the influence of perceptual factors on AV integration capacity, in that congruency at 200 ms SOA failed to significantly impact performance and failed to raise capacity estimates above 1. Such a constraint is also consistent with the temporal frequency limit reported by Holcombe and Chen (2013), whereby in order to reliably track two visual stimuli, the rate of change between them needed to be minimally 250 ms. Therefore, if the capacity of visual object tracking cannot exceed 1 at 200 ms then it similarly seems unlikely that the capacity of audio-visual integration could exceed 1. Such ideas are also echoed in the behavioural observations of Van der Burg et al. (2013) where a slowing in SOA leads to a reduction in the number of incorrect audio-visual bindings, and in the neural data of Wilbiks & Dyson (2016) where visual N1 amplitude was sensitive to the number of changing locations per frame during slow but not fast rates of visual stimulus presentation, potentially representing poor quality sensory information entering working memory under fast (e.g., 200 ms) SOA conditions.

A further reason why cross-modal correspondences may have had an effect at 700 ms SOA but not at 200 ms SOA is that congruency between visual polarity change and auditory frequency was only revealed on the single, critical, frame. In this respect, the perceptual manipulation intended to enhance AV capacity was a transient rather than sustained one. Therefore, to examine the impact of sustained (and assumedly, stronger) perceptual effects on the facilitation of audio-visual integration capacity, we examined the role of chunking in Experiment 2. By attempting to consolidate multiple visual

locations throughout the entire trial (chunking) rather than just during the critical frame (cross-modal congruency), this sustained effect may have a more dramatic impact on the capacity of audio-visual integration.

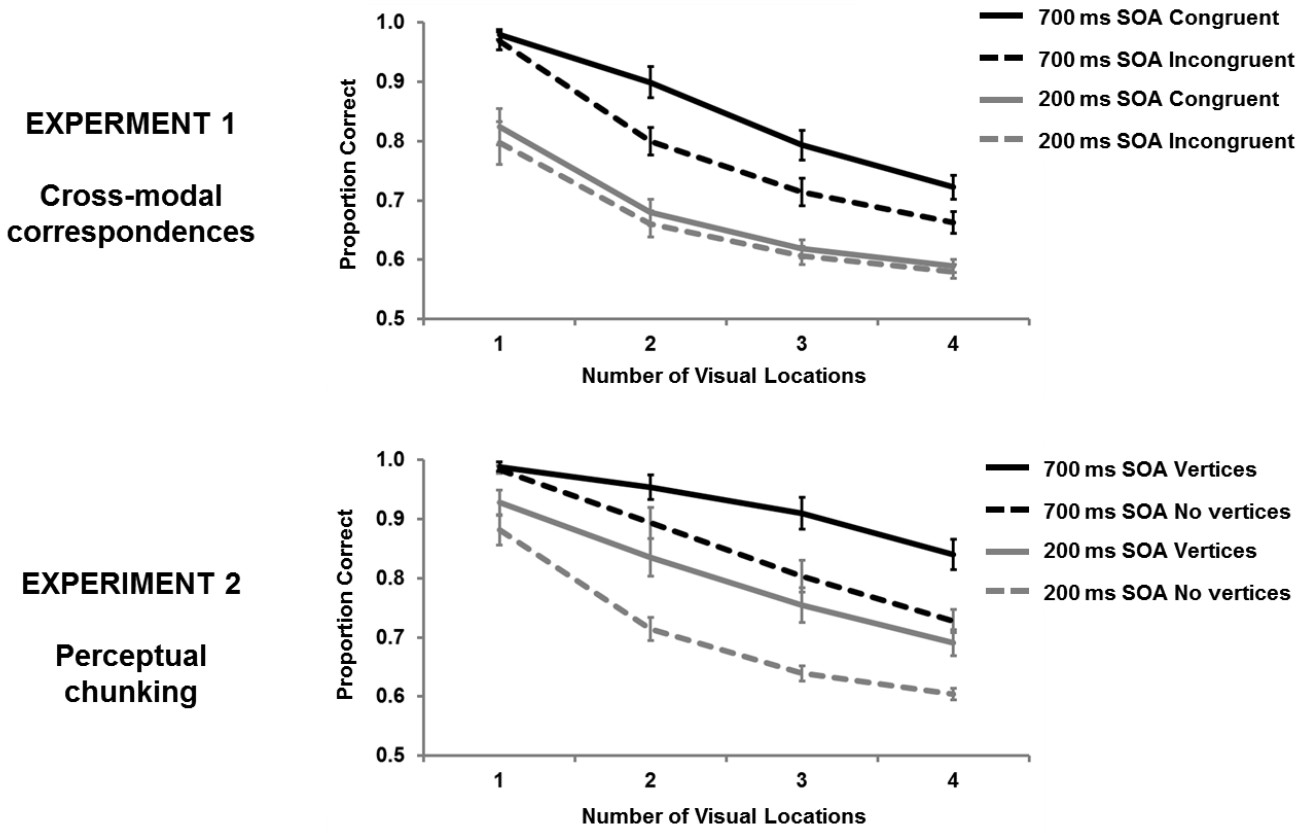

Figure 3. Proportion correct responding as a function of SOA (200 ms (grey) or 700 ms (black)), crossmodal (in)congruency between colour switch and tone pitch at the critical trial (Experiment 1), and, perceptual chunking mechanisms via the presence or absence of vertices connecting visual locations (Experiment 2). Error bars represent standard error

However, the facilitation of audio-visual integration by cross-modal congruency under slower presentation speeds provides support for the argument that audio-visual integration is present at these slower speeds, which is contrary to some earlier research (e.g. Van der Burg et al., 2013).

Experiment 2

In terms of the uni-modal literature, working memory span has been shown to functionally increase by means of a technique called chunking. First described by Miller (1956), this technique involves combining multiple items to be held in working memory into more complex, but less numerous items, allowing for a greater amount of information to be maintained in working memory. For example, in the learning of language, chunking is implemented via both bottom-up (based on statistical regularities) and top-down (based on familiarity with words) routes, yielding more efficient reading (Jones, Gobet, & Pine, 2007). While chunking has traditionally been discussed in terms of working memory, it has also been shown to be an effective perceptual aid. For example, Gobet and Simon (1998) considered expert chess players' perception of chess positions and found that, while non-experts perceive positions of each piece independently and then build a concept of the game situation, expert players perceive the chessboard as a chunk, a single situation including all piece positions.

Additional evidence from non-expert participants also show that perceptual chunking improves performance. Gmeindl, Walsh, and Courtney (2011) presented participants with a display of scattered grey squares, with some designated targets (via a black outline) and others as distractors (no outline). Participants were asked to indicate

targets either by touching all targets or typing the locations on a keyboard. Their results indicated that people performed better when engaging in the spatial task of touching rather than typing, and this effect was increased as a function of the nearness of the targets to one another in the display. The authors propose that this was evidence for the use of perceptual chunking, as participants were better able to perform the task when it was a spatial one, and when targets can be grouped. Sargent, Dopkins, Philbeck, and Chichka (2010) provide similar evidence for perceptual chunking as a technique. Here, participants were exposed to targets arranged 360-degrees around them in a room. When attempting to identify them, performance was improved if targets were closer to one another, within an arrangement that was seen multiple times within the experiment, and, when that arrangement could be mapped onto a common object. For example, if two balloons were being employed as targets, participants were better able to identify them better if they were attached to two corners of a blackboard than if they were attached to two disparate locations on the wall, even if the balloons were the same distance away from one another in both instances. This final explanation is most pertinent to the current research – using an object to chunk together disparate target locations may allow for more information to be successfully tracked and ultimately bound to the auditory modality.

In Experiment 2, effects of perceptual chunking on capacity of audio-visual integration will be examined by- essentially- connecting the dots for the participant. By using connected vertices overlaid on the dots that change at each frame, participants should be able to perceive one, complex object rather than a greater number of simple

objects. Since the presentation of vertices will appear on all frames (and not just the critical frame), this should represent a more sustained and stronger perceptual effect than Experiment 1. Like cross-modal congruency, perceptual chunking should facilitate the binding of auditory and visual information such that the functional capacity of audio-visual integration may exceed the putative upper bound of 1.

Method

29 new participants took part in the study. Data were trimmed as in Experiment 1, with the final sample consisting of 24 participants with an average age of 22.0, with 20 females and 24 right handed individuals. Again, using the observed effect size of η_p^2 = .177 from Wilbiks & Dyson (2016; Experiment 4), a minimum sample size of 20 participants was required assuming a similar effect size, with $\alpha = .05$ and Power $(1 - \beta) = .80$. As in Experiment 1, our exclusion procedure was recursive, and as such it was not always possible to stop on a specific number, hence the slightly larger sample size in Experiment 2. Experiment 2 was identical to Experiment 1 apart from two changes. First, only the low (300 Hz) tone was used. Second, on half of the trials, in addition to a fixed number of dots changing at each alternation, vertices were presented in a mid-grey colour (100, 100, 100) at those same locations in the form of: a dot with a diagonal slash on it (when 1 dot changed), a line (2-dot change), a triangle (3-dot change), or a quadrilateral (4-dot change; see Figure 1b for a schematic of a 2-dot change trial with vertices). While we acknowledge the difference between a single stimulus including a slash and multiple connected stimuli, we included additional visual information on the single stimulus in order to maintain parity across number of locations in the vertices condition. This design

feature also helped to maintain a clear visual difference between the vertices and no vertices conditions even when only a single location was changing. The number of locations / number of vertices to be tracked and SOA were manipulated within blocks, while vertices (present, absent) was manipulated across blocks. Participants completed a practice block of 16 trials and 4 experimental blocks (48 trials in each) of both the vertex and no-vertex condition, for a total of 384 experimental trials, with condition order counterbalanced across individuals.

Results

Capacity measures (*K*) were calculated as in Experiment 1, with goodness of fit confirmed by low RMSEs ranging from 0.0001 – 0.1581. *K* was subjected to a 2 x 2 within-participants ANOVA with factors of vertices (absent, present) and SOA (200 ms, 700 ms). The resultant data are shown in the lower panel of Figure 2. A main effect of vertices: $F(1,23) = 59.34$, MSE = 0.262, $p < .001$, $\eta_p^2 = .721$, and SOA: $F(1,23) = 106.07$, MSE = 0.272, $p < .001$, $\eta_p^2 = .822$, were shown, in the absence of a significant interaction: $F(1,23) = 2.05$, MSE = 0.149, $p = .165$, $\eta_p^2 = .082$. Thus, capacity estimates were larger during slow relative to fast stimulus presentation as it was Experiment 1, the presence of vertices also increased *K* relative to their absence, and, the influence of vertices was equivalent between 700 (1.82 and 2.74) and 200 (0.84 and 1.53) ms SOA conditions. In comparing group estimates of audio-visual capacity against the critical value of 1, for the 700 ms SOA, *K* was significantly greater than one with ($t[23] = 8.35$, $p < .001$) and without vertices ($t[23] = 5.15$, $p < .001$). Capacity remained significantly less

than 1 when vertices were absent during the 200 ms SOA (t[23] = -2.195, p = .038) but was significantly greater than 1 in the presence of vertices (t[23] = 3.04, p = .006).

As in Experiment 1, the proportion of correct responses for each combination of SOA, number of locations changing, and level of chunking was analyzed (means and standard errors displayed in Figure 3). The data were submitted to a 2 (SOA: 200 ms, 700 ms) x 2 (vertices: absent, present) x 4 (number of locations: 1, 2, 3, 4) repeated measures ANOVA. We found expected significant effects of SOA, ($F(1,23) = 125.79$, MSE = 0.013, $p < .001$, $\eta_p^2 = .845$), vertices, ($F(1,23) = 46.17$, MSE = 0.014, $p < .001$, $\eta_p^2 = .667$), and number of locations, ($F(3. 69) = 273.24$, MSE = 0.003, $p < .001$, $\eta_p^2 = .922$). There were also significant interactions of vertices x number ($F(3, 69) = 23.27$, MSE = .002, $p < .001$, $\eta_p^2 = .503$) and SOA x number ($F(3, 69) = 9.56$, MSE = .003, $p < .001$, $\eta_p^2 = .294$), but not vertices x SOA ($F(1, 23) = 0.72$, MSE = .015, $p = .404$, $\eta_p^2 = .030$), which are explained by a significant three-way interaction, ($F(3, 69) = 4.53$, MSE = 0.002, $p = .006$, $\eta_p^2 = .165$). Post hoc comparisons using Tukey's HSD ($p < .05$) revealed that response accuracy was facilitated by the presence of vertices in all conditions except for 700 ms SOA with 1 location changing. However, for the first time we see a facilitatory effect at 200 ms SOA, at all numbers of locations changing.

Discussion

The perceptual chunking of multiple, independent, dot polarity changes into a single complex object enabled the capacity of audio-visual integration to exceed the putative limit of 1 (Van der Burg et al., 2013), this time during both slow and fast rates of presentations. The data from Experiment 2 are particularly important as they rule out an

alternative account suggesting that estimates of capacity only exceed 1 when delivery rates are slow, and as such the perceptual conditions observed at 700 ms SOA do not represent 'true' audio-visual integration. Our observation of $K > 1$ at both 200 and 700 ms SOA instead reinforce the idea that chunking is an effective strategy for effectively increasing perceptual span, in both uni-modal (Gmeindl, Walsh, & Courtney, 2011; Gilbert, Boucher & Jemel, 2014; Sargent et al., 2010; van Meeuwen, Jarodzka, Brand-Gruwel, Kirschner, de Bock, & van Merriënboer, 2014) and also– now– in multi-modal contexts.

Our previous work suggested that, at a neural level, there was some difficulty in distinguishing between the number of polarity changes that were occurring *prior to* the critical audio-visual trial (Wilbiks & Dyson, 2016, Experiment 4). This raised the possibility that the *K* limit of 1 observed at fast SOA was not a limit of AV integration but rather a limit in the ability to parse, track and update potentially multiple visual locations. We introduced the vertices manipulation in Experiment 2 as a form of perceptual grouping to increase the likelihood that participants had more accurate information about the change (or non-change) of multiple locations before the moment of audio-visual integration; the addition of the vertices did not provide any additional information as to the validity (or invalidity) of to-be-probed location itself.

In considering the results from Experiment 2, we are reminded of a debate in the literature about the nature of visual working memory span – namely, is it measured strictly by a number of objects, or rather by a combination of number of objects and complexity of those objects? Awh, Barton, and Vogel (2007) propose that the capacity of

visual working memory is around 4 items, and that this limit is not affected by the level of complexity of items. Alvarez and Cavanagh (2004), on the other hand, provide evidence that the capacity of visual working memory is limited by both the number of objects, and the relative complexity of those objects. In Experiment 2, the capacity of audio-visual integration at 200 ms improves to the point where it is greater than one. In cases where capacity was closer to two however, it is possible that participants were not tracking two dots, but rather the orientation of a line connecting those dots. Looking at the data from this perspective suggests that the true numerical capacity of integration is still one item at 200 ms, but that the *functional* capacity can be increased by means of perceptual chunking. This accords with Awh, Barton, & Vogel's (2007) conceptualization of working memory, wherein the same number of objects can be held in visual working memory (approximately 4), regardless of complexity.

We can also draw on research into feature binding to inform the current research on perceptual chunking. The comparison between the configural hypothesis and the elemental hypothesis in feature binding (Moeller, Frings, & Pfister, 2016; Moeller, Pfister, Kunde, & Frings, 2016) can be used to explain the results of Experiment 2, and in some ways this debate seems analogous to the comparison between objects and complexity in working memory. According to this perspective, the configural hypothesis states that associations are formed between entire stimuli and their respective responses, while the elemental hypothesis states that features within stimuli can be bound to each other and to responses independently. This dichotomy seems to be a parallel with

considering multiple locations that are connected as a single, complex stimulus with complexity that does not modulate capacity (i.e. Awh, Barton, & Vogel, 2007) or as multiple simpler stimuli, with increasing complexity leading to reduced numerical capacity (i.e. Alvarez & Cavanagh, 2004). In the current Experiment 2, we find support for numerical capacity for audio-visual integration as a factor independent from stimulus complexity, such that at a 200 ms SOA, only 1 visual object can be integrated with an auditory stimulus, but that this object can be either simple (a dot) or complex (a line or polygon).

Experiment 3

The results from Experiments 1 and 2 show that the estimated capacity of audio-visual integration can exceed 1 by using transient congruence between the two modalities on the critical frame (Experiment 1) and by utilizing sustained perceptual chunking of visual information during the trial (Experiment 2). However, concerns may be raised in our repeated observations of demonstrating $K > 1$ in slow (700 ms) relative to fast (200 ms) SOA conditions. Previous reviewers have made the suggestion that there is a qualitative difference between the two rates of presentation – that an SOA of 200 ms or less represents "true" audio-visual integration, while at 700 ms SOA the task can be completed on the basis of visual information alone. First, if this perspective is to be accepted, one would expect capacity at 700 ms to approach that of visual short term memory, which has been shown to be between 3 and 4 items (Cowan, 2001), and this is not the case. Second, a visual-only control condition in which the auditory cue for the critical frame was replaced by a visual cue (Wilbiks & Dyson, 2016, Experiment 5; after

van der Burg et al., 2013) showed no facilitation of capacity at slow or fast SOAs, and as such capacity is promoted via the integration of audio-visual information in a way that the integration of visual-visual information does not. Third, in our current Experiment 2, we demonstrated an additive effect of vertices on both the 200 ms and 700 ms SOA conditions and that- at a group level- *K* exceeded 1 during both fast and slow visual presentation. Nevertheless, to further address the idea whether manipulations of SOA represent qualitative or quantitative shifts in multi-modal processing we will examine the effects of training across three SOAs in Experiment 3.

Training has been shown to have an influence on multi-modal integration, often evidenced through recalibration of the temporal window of integration. Fujisaki, Shimojo, Kashino, and Nishida (2004) presented participants with an auditory and a visual stimulus and asked them to judge whether the two stimuli were presented simultaneously or not. They manipulated the lag between the visual and auditory stimuli systematically, and in doing so induced a recalibration of participants' point of subjective simultaneity such that it shifted towards the manipulated lag. That is to say, presenting a large number of trials where the visual stimulus preceded the auditory stimulus by, on average, 100 ms led participants to perceive simultaneity. Heron, Roach, Hanson, McGraw, and Whitaker (2012) expand on this idea by showing that while recalibration within a set of stimulus presentations tends to be 'attractive' (that is, move towards the preset lag prescribed by the experiment), there can also be 'repulsive' aftereffects, wherein the newly calibrated system shifts away from the manipulated lag once the manipulation is over. Work in our own laboratory (Wilbiks & Dyson, 2013b) found

evidence for the repulsive aftereffects described by Heron et al. (2012) when participants made decisions about which of two visual sources was more likely to have generated a single auditory event.

We extend these ideas by examining whether the *capacity* of audio-visual integration can also be increased through training. We expect that participants will show an increase in audio-visual integration capacity, specifically at an intermediate SOA that is used for training (450 ms), but that this may also generalize to other SOAs that are included in the experiment (200 and 700 ms; c.f. transfer effects; Brehmer, Westerberg & Backman, 2012). If there is a qualitative different between the mechanisms of audio-visual integration at 200 ms relative to 700 ms SOA, then we predict that the amount of transfer from the intermediate, trained SOA to fast and slow, untrained SOAs should differ. Additionally, if qualitative differences exist between audio-visual integrative processes at fast and slow rates of presentation, then this should also yield a non-linear trend across the three SOAs both pre- and post-training.

Method

Participants. A minimum sample size of 17 participants was required assuming a previous effect size of $\eta_p^2 = .177$, with $\alpha = .05$ and power $(1 - \beta) = .80$ and the requirement to evaluate a 2x3 ANOVA. We recruited 36 participants, but 10 of them failed to attend both testing sessions, or had a computer error during recording, meaning we were left with 26 complete and viable data sets. All participants were recruited from an undergraduate research participant pool, and were compensated with partial class credit. Five participants were removed as a result of performing within the 95% CI for

chance responding across all conditions, so the final sample consisted of 21 participants, with a mean age of 20.2, including 17 right handed individuals and 16 females. All participants were recruited from an undergraduate research participant pool, and were compensated with partial class credit. Five participants were removed as a result of performing within the 95% CI for chance responding across all conditions, so the final sample consisted of 21 participants, with a mean age of 20.2, including 17 right handed individuals and 16 females. None of the participants took part in any of the previous experiments in this series. Each participant signed up for two 1-hour testing sessions, which were always scheduled for consecutive days. On Day 1, the participant initially completed a Test Session followed by a Training Session. On Day 2, the participants completed a Training Session, followed by a Test Session (see below for details).

Design and Procedure. Stimuli and stimulus presentation were identical to Experiment 1, except that there was no manipulation of congruency, and there was an additional SOA of 450 ms. Each block orthogonally varied the SOAs (200, 450, 700 ms), the validity of the stimulus (valid, invalid), and the number of visual stimuli changing (1, 2, 3, or 4). Each block consisted of trials with the orthogonal combinations of factors, and participants completed 4 test blocks in each Test Session, with each block comprising 48 trials. The training block consisted of only a single SOA (450 ms), but still contained the combination of validity and number of stimuli changing as before. Each training block contained 3 repetitions of the 8 combinations of validity and number of stimuli, making for 24 trials in each block. Participants completed 10 training blocks in each Training Session. Participants were offered the chance to complete a practice block

consisting of 12 randomly chosen trials before beginning their first test block and their first training block of each session. Trial order was randomized in practice and in experimental trials and validity was collapsed for analysis purposes.

Results

Model fitting was conducted as in the previous experiments, with successful model fit confirmed by average RMSE of 0.068 (range: 0.005 - 0.154), 0.037 (range: 0.001 - 0.113), and 0.041 (range: 0.001 - 0.162) for 200, 450, and 700 ms conditions, respectively. Capacity estimates for each set of conditions are displayed graphically in Figure 4.

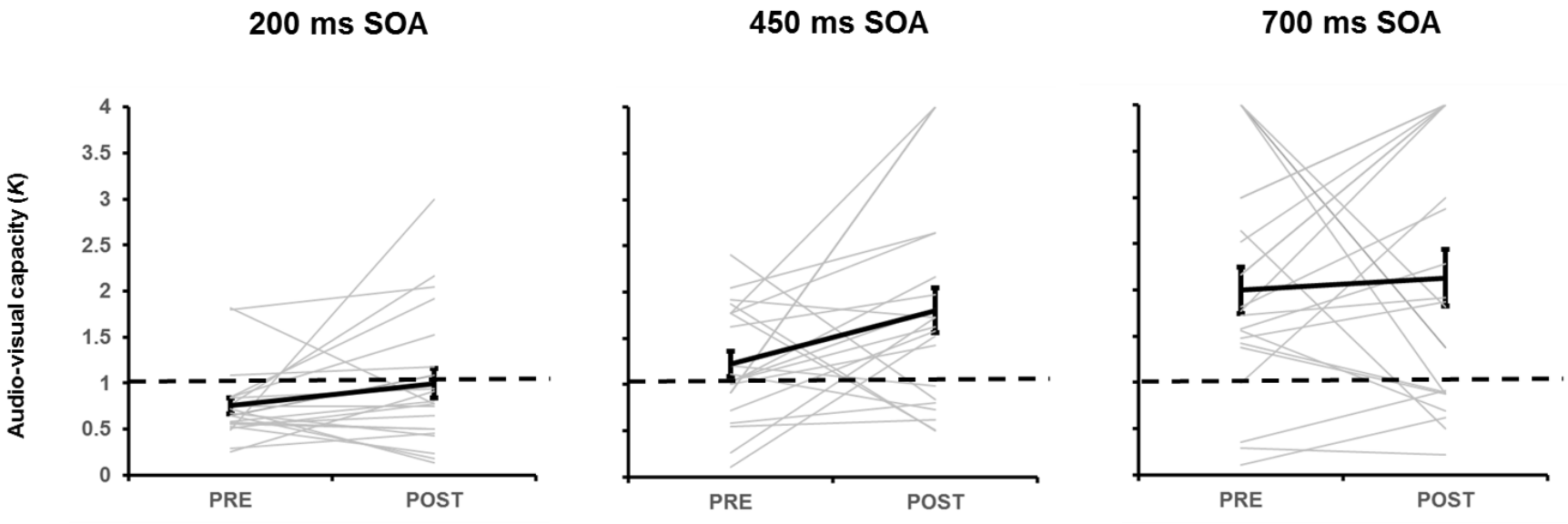

Figure 4. Audio-visual integration capacity (K) as a function of SOA (200, 450, or 700 ms) and training (pre- or post-training), as in Experiment 3. Grey lines represent individual participant data, whereas black lines represent group average data (error bars represent standard error).

Data from test blocks before and after training sessions were submitted to a 2 (training: pre, post) x 3 (SOA: 200, 450, 700 ms) repeated measures ANOVA. The main effect of training was not significant, $F(1, 20) = 2.784$, MSE = 1.153, $p = .111$, $\eta_p^2 = .122$. A main effect of SOA was significant, $F(2, 40) = 31.616$, MSE = .471, $p < .001$, $\eta_p^2 = .613$, with a significant increase in capacity from 200 to 450, as well as from 450 to 700 ms (Tukey's HSD, $p < .05$). The interaction between training and SOA was not significant, $F(2, 40) = .991$, MSE = .577, $p = .380$, $\eta_p^2 = .047$. Comparisons between pre- and post-training estimates of *K* were not significant for 200 ($t[20] = 1.498$, $p = .150$) or 700 ($t[20] = .386$, $p = .704$) ms, but were significant for 450 ms ($t[20] = 2.111$, $p = .048$), suggestive of criterion but not transfer effects for audio-visual training in Experiment 3. *K* exceeded 1 in the 700 ms SOA condition both pre- and post-training, and, the 450 ms SOA condition post-training (see Table 1).

Table 1 – Capacity for each combination of SOA and training with single sample t-tests against test value of 1.

Measure	*K*	*t*(20)	*p*
Pre-training / 200 ms	.751	-2.87	.010
Pre-training / 450 ms	1.219	1.62	.121
Pre-training / 700 ms	1.997	3.97	.001
Post-training / 200 ms	.996	-0.26	.979
Post-training / 450 ms	1.800	3.284	.004
Post-training / 700 ms	2.129	3.662	.002

In order to ascertain whether effects of training are stronger for certain stimulus combinations than others, we examined the proportion of correct responses for each stage of the experiment, SOA, and number of locations changing (means and standard errors displayed in Figure 5). The data were submitted to a 2 (training: pre-training, post-training) x 3 (SOA: 200, 450, 700 ms) x 4 (number of locations: 1, 2, 3, 4) repeated measures ANOVA. We found expected significant effects of SOA, ($F(1,20) = 34.42$, MSE = 0.026, $p < .001$, $\eta_p^2 = .632$) and number of locations, ($F(3, 60) = 264.22$, MSE = 0.004, $p < .001$, $\eta_p^2 = .930$), but not training, ($F(1, 20) = 1.97$, MSE = 0.076, $p = .176$, $\eta_p^2 = .090$). In this case, there was only one significant interaction, which was between SOA

and number of locations changing, ($F(6, 120) = 3.04$, MSE = 0.004, $p = .008$, $\eta_p^2 = .132$). Post hoc comparisons using Tukey's HSD ($p < .05$) showed that response accuracy was significantly affected by SOA at all numbers of locations changing – higher accuracy was observed at 700 ms than 450 ms, and at 450 ms than at 200 ms, at all numbers of locations changing.

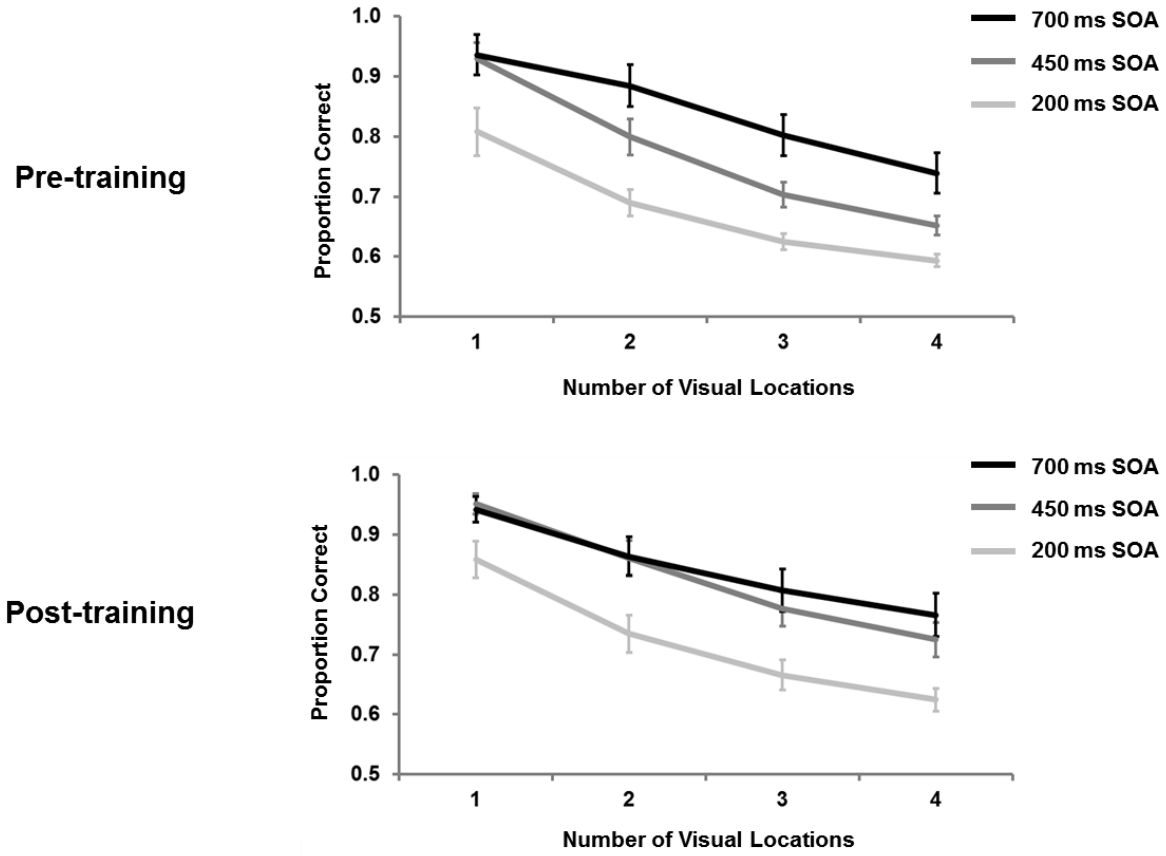

Figure 5. Proportion correct responding as a function of SOA (200, 450, or 700 ms) and training (pre- or post-training), as in Experiment 3. Error bars represent standard error.

Finally, to test for qualitative differences in audio-visual integration mechanisms during fast and slow rates of presentation, both pre- and post-training data across the three SOAs were submitted to trend analysis. Both pre- and post-training data revealed linear trends (pre-training: $F(1,20) = 26.264$, MSE = .622, $p < .001$, $\eta_p^2 = .568$; post-training: $F(1,20) = 20.445$, MSE = .660, $p < .001$, $\eta_p^2 = .505$), in the absence of quadratic trends (pre-training: $F(1,20) = 1.851$, MSE = .181, $p = .189$, $\eta_p^2 = .085$; post-training: $F(1,20) = 1.244$, MSE = .633, $p = .278$, $\eta_p^2 = .059$). Therefore, the data are in support of quantitative rather than qualitative differences between the mechanisms evoked during relatively fast and relatively slow audio-visual integration.

Discussion

The findings from Experiment 3 support criterion training effects (Brehmer et al., 2012) in enhancing the capacity of audio-visual integration. When participants completed 480 training trials, over two days, with an SOA of 450 ms, their capacity of audio-visual integration at 450 ms was significantly increased. Therefore, not only can task parameters such as SOA, temporal predictability, and the degree of proactive interference (Wilbiks & Dyson, 2016) modulate the capacity of audio-visual capacity, but so can external perceptual factors (Experiments 1 and 2) and internal recalibration as a result of training (Experiment 3). Importantly, there is no data in Experiment 3 that point to qualitative differences in the processes associated with audio-visual integration as a function of SOA. First, we failed to find an interaction between training and SOA and follow-up analyses suggested that there were no transfer effects to either slow (700 ms) or fast (200 ms) SOA conditions (although naturally we need to be cautious about the

interpretation of a null result). Second, our trend analyses both pre- and post-training provided evidence for only a linear relationship between SOA and *K*. Experiment 4 sought to derive more conclusive evidence regarding the linearity of *K* as a function of SOA.

Experiment 4

If the mechanisms of the current task qualitatively shift from genuine audio-visual integration at short SOAs to visual-only processing at long SOAs, then we would expect to see a non-linear trend in the data, as *K* switches from the audio-visual integration capacity of 1 to the much larger capacity of 3-4 estimated for VSTM. An analogy can be drawn here between the comparison of qualitative and quantitative differences in multi-modal binding and previous research into visual and informational persistence in visual short term memory. Early research in this field provided evidence that briefly presented visual stimuli led to visual persistence – it remained perceptually visible for a brief time after it was no longer being presented (cf. Sperling, 1960; Neisser, 1967). Later research, however, revealed that visual persistence only existed at faster presentation speeds, while slower presentation speeds allowed for information to persist in short term memory, but without an iconic visual representation (Di Lollo & Wilson, 1978; Di Lollo, 1980; Irwin & Yeomans, 1986). As a result, data generated from visual vs. information persistence take a non-linear form when performance is studied across a number of SOAs (e.g. Loftus & Irwin, 1998, Figure 2, left panel). Similarly, if performance in the current task relies upon qualitatively difference processes between fast and slow SOA we too would expect to see a non-linearity in the functional capacity of *K*. In Experiment 4, we

extracted SOA data from a study in a forthcoming experimental series (Wilbiks, Rioux & Dyson, in preparation) in which *K* was evaluated from 100 ms to 600 ms.

Method

32 participants were recruited from Introductory Psychology courses at Mount Allison University and were compensated with partial course credit in exchange for one hour of participation. After participant exclusion, Experiment 4 yielded 26 participants who had a mean age of 19.8 years (SD = 1.2), included 25 females, and were all right handed. All stimulus details, design and procedure were the same as in Experiment 2, apart from the replacement of the vertices present condition with a condition in which the lines that were drawn to connect polarity changing locations were the same colour as the background (128, 128, 128). SOA was now also examined in 100 ms intervals from 100 ms SOA to 600 ms SOA.

Results

Model fitting was conducted in the same way as in the previous experiments (RMSE range from 0.0001 - .2465). A trend analysis was conducted on the SOA data (see Figure 6), and a significant linear trend was found ($F(1,25) = 42.633$, MSE = .962, $p < .001$, $\eta_p^2 = .630$), in the absence of a quadratic trend ($F(1,25) = .402$, MSE = .253, $p = .532$, $\eta_p^2 = .016$), or any higher order trends (all $F < 1.79$, $p > .194$)[1].

[1]Analysis of the condition without illusory contours also revealed a significant linear trend ($F(1,25) = 30.229$, MSE = .354, $p < .001$, $\eta_p^2 = .547$), in the absence of any higher order trends (all $F < 12.72$, $p > .109$).

Discussion

The data from Experiment 4 provide evidence of quantitative rather than qualitative changes in *K* as a function of SOA, as a result of a significant linear trend in the absence of any higher order effects. Although we acknowledge that as-yet unspecified interactions between differentially weighted modules involved in AV integration might yet account for the observed pattern, we believe that the application of the same mechanisms during AV integration across both slower and faster SOAs is the most parsimonious explanation of a linear trend in the data. We further took this opportunity to consolidate a number of experiments evaluating the capacity of audio-visual integration using variants of the pip-pop paradigm (van der Burg, Olivers, Bronkhorst & Theewues, 2008), in which SOA was manipulated (see Figure 6). Here, we replot the data from the current four experiments, Experiment 4 from Wilbiks & Dyson (2016; upon which the current series was based), and, Experiment 1c from van der Burg et al (2013). Two observations are particularly salient. First, despite differences in intercept value, our own data is relatively consistent in the slope of *K* as a function of SOA. Therefore, we continue to find no evidence in our data that the mechanisms of the task qualitatively shift from genuine audio-visual integration at short SOAs (a limit of 1) to visual-only processing at long SOAs (a non-linear shift in limit to 3-4). Second, the SOA manipulation of just 50 ms (from 150 ms to 200 ms) led to a significant increase in *K* from 0.79 to 1.05 in van der Burg et al., 2013 (Experiment 1c). It is clear that if one were to extrapolate (albeit cautiously) their data to an SOA of 250 ms, then *K* would be estimated at 1.31 and thus exceed 1 at a group level. The implications of this second point

are either that a) true audio-visual integration mechanisms can only be observed when the visual presentation rate is 200 ms or less (whereby these mechanisms are also reliably disrupted by increasing the number of mis-binding between vision and audition as a result of visual presentation speed; van der Burg et al., 2013, p. 348), or, b) that the capacity of audio-visual integration is not limited to 1 (*contra* van der Burg et al., 2013).

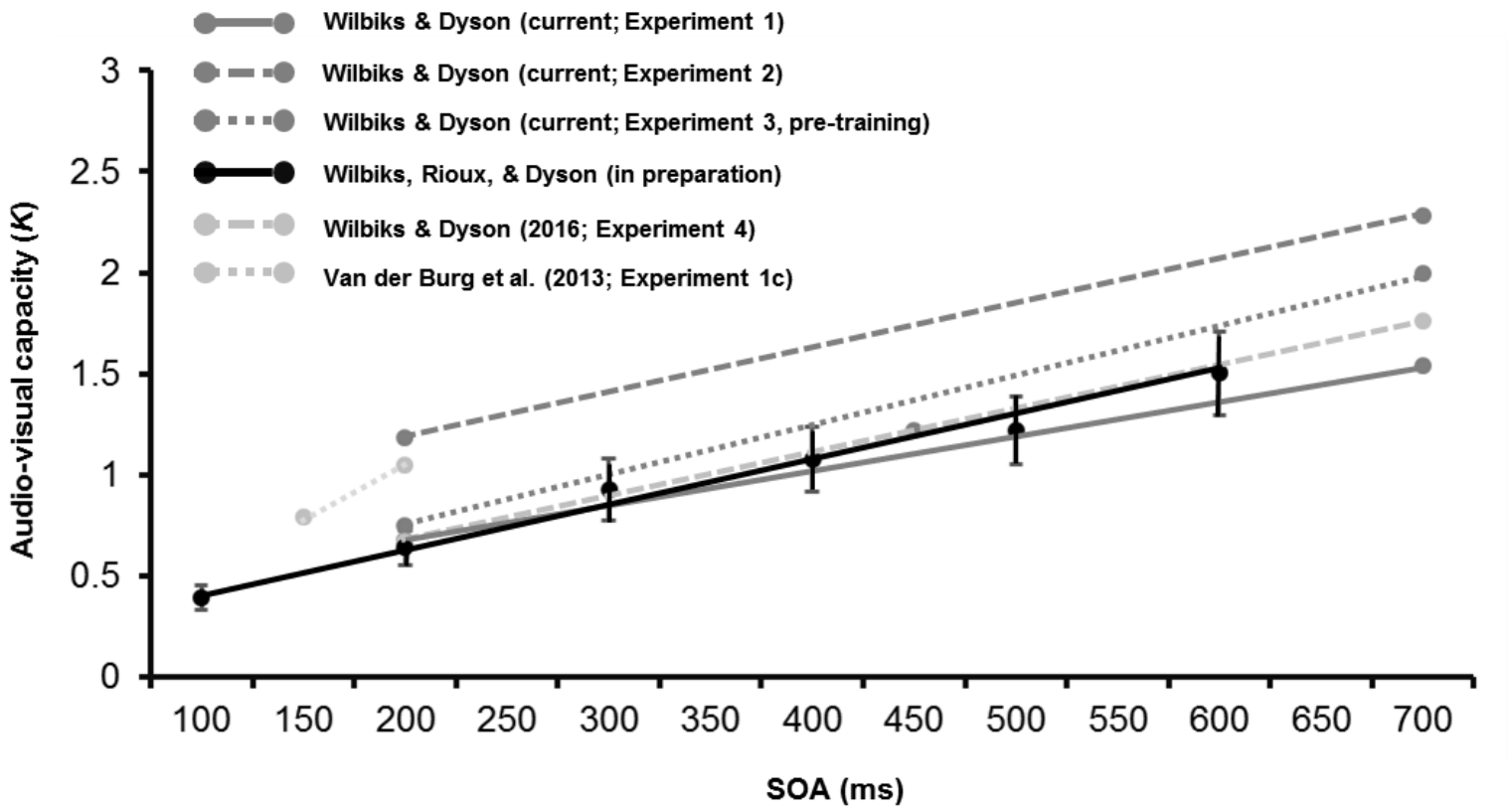

Figure 6. Capacity estimates from six experiments (including the present Experiment 4) indicating similar, linear trends across increasing SOAs.

General Discussion

Across four experiments, we focused on a variety of internal and external factors that might facilitate multi-modal processing, in order to challenge the presumption that the upper-limit of audio-visual integration is 1 (after Van der Burg et al., 2013). Experiment 1 showed that using a transient brightness-pitch cross-modal correspondence between vision and audition (e.g., Marks, 1987; Parise & Spence, 2009) produced estimates of audio-visual capacity greater than 1 during slow rates of visual presentation (where SOA = 700 ms). In contrast, Experiment 2 used more sustained perceptual chunking (e.g., Gobet et al., 2001) via the use of vertices and showed that *K* was significantly greater than 1 during both slow (SOA = 700 ms) and fast (SOA = 200 ms) delivery. Experiment 3 showed that at an intermediate SOA (450 ms) K exceeded 1 post-training, although no transfer effects were observed. Experiment 4 provided evidence for the linearity of *K* as a function of SOA, ruling out the suggestion that the way the current task is completed is qualitatively different between slow and fast speeds of processing. Finally, we consolidated the current research into the capacity of audio-visual integration as a function of SOA by showing that our own previous research has produced similar *K* slopes indicative of a continuum of performance across SOA, and that other research is likely to have observed K > 1 by slowing their paradigm down by a further 50 ms. In short, we find numerous examples where the capacity of audio-visual integration is not limited to 1.

Given that the nature of cognition is rarely fixed, one long-term goal of the field should be to delineate the conditions under which processes such as those represented by audio-visual integration capacity operate in a number of observable ways. In contrast to previous work (Van der Burg et al., 2013; Olivers et al., 2016), we present an alternative view of audio-visual integration capacity, one that is malleable and can be influenced by environmental and individual demands. The flexibility of audio-visual integration capacity is

entirely consistent with similar malleability previously observed in both uni-modal (Drew, Horowitz & Vogel, 2013) and multi-modal (Fujisaki, Shimojo, Kashino & Nishida, 2004) literatures. In examining both the individual, as well as group average data in Figures 2 and 4, there is also a great deal of inter-individual difference in AV capacity (*K*) and participants appear idiosyncratically impacted by the various manipulations deployed. This seems particularly true for the training protocol in Experiment 3 and suggested that, for some individuals, training might have led to a deterioration of near-ceiling performance at untrained SOAs (i.e., 700 ms). Given the temporal separation between pre- and post-training sessions, variation in testing time and hence alertness may have also given rise to these changes. Clearly, a more comprehensive consideration of the cognitive state of the participant at the time of testing in terms of factors like attention and working memory capacity will help to explain why performance is so variable across individuals.

We find evidence in support of a distributed attention perspective (Zhang & Luck, 2008; Huang, 2010), which is not in alignment with Olivers et al. (2016) single source hypothesis, in that participants are able to use both transient (Experiment 1) and sustained (Experiment 2) features to increase capacity. If, as Olivers et al. (2016) propose, individuals are only attending to specific stimuli as they are presented and integrated, a transient feature would not be able to serve a facilitative role, as this would only be possible if they were able to attend to the full visual display in the first place. The present research also anticipates a question asked by Olivers et al. (2016) – to wit, ”Could such grouped events [e.g. visual stimuli grouped by proximity, color, or shape] count as a single event for mechanisms of audio-visual integration?“ (p. 2122). We find that grouping via perceptual chunking can lead to an increase in capacity. It is clear, then, that ‘grouped events’ of this type can lead to an increase in capacity – although the argument of whether they represent a single event remains a theoretical debate, pending future research.

While it is possible that there is some overarching maximum limit on the capacity of audio-visual integration, such a limit was not observed in this experimental series, nor was it observed in the five experiments in Wilbiks and Dyson (2016). As such we are accumulating a growing corpus of evidence in favour of a flexible capacity for audio-visual integration. Future research will provide definitive answers as to whether a maximum limit does exist, as well as to many of the questions asked above. What can be stated with certainty at this point is that capacity varies based on stimulus factors and individual training effects, and that in certain contexts audio-visual integration capacity can be raised above the putative limit of 1.

References

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*(2), 106-111.

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science, 18*(7), 622-628.

Brehmer, Y., Westerberg, H., & Backman, L. (2012). Working-memory training in younger and older adults: training gains, transfer, and maintentance. *Frontiers in Human Neuroscience, 6*(63), 1-7.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity [Target article and commentaries]. *Behavioral and Brain Sciences, 24*, 87-185.

Di Lollo, V. (1980). Temporal integration in visual memory. *Journal of Experimental Psychology: General, 109*(1), 75.

Di Lollo, V., & Wilson, A. E. (1978). Iconic persistence and perceptual moment as determinants of temporal integration in vision. *Vision Research, 18*(12), 1607-1610.

Drew, T., Horowitz, T. S., & Vogel, E. K. (2013). Swapping or dropping? Electrophysiological measures of difficulty during multiple object tracking. *Cognition, 126*(2), 213-223.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*, 6.

Faul, F., Erdfelder, E., Lang, A-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fiebelkorn, I. C., Foxe, J. J., & Molholm, S. (2010). Dual mechanisms for the cross-sensory spread of attention: How much do learned associations matter? *Cerebral Cortex, 20*(1), 109-120.

Frings, C., Rothermund, K., & Wentura, D. (2007). Distractor repetitions retrieve previous responses to targets. *Quarterly Journal of Experimental Psychology, 60*, 1367-1377.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7*(7), 773-778.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics, 68*(7), 1191-1203.

Gilbert, A. C., Boucher, V. J., & Jemel, B. (2014). Perceptual chunking and its effect on memory in speech processing: ERP and behavioral evidence. *Frontiers in Psychology, 5.*

Gmeindl, L., Walsh, M., & Courtney, S. M. (2011). Binding serial order to representations in working memory: a spatial/verbal dissociation. *Memory & Cognition, 39*(1), 37-46.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*(6), 236-243.

Gobet, F., & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory, 6*(3), 225-255.

Heron, J., Roach, N. W., Hanson, J. V., McGraw, P. V., & Whitaker, D. (2012). Audiovisual time perception is spatially specific. *Experimental Brain Research, 218*, 477-485.

Heron, J., Whitaker, D., McGraw, P. V., & Horoshenkov, K. V. (2007). Adaptation minimizes distance-related audiovisual delays. *Journal of Vision, 7*(13), 5.

Holcombe, A. O., & Chen, W. Y. (2013). Splitting attention reduces temporal resolution from 7 Hz for tracking one object to <3 Hz when tracking three. *Journal of Vision, 13*(1), 1-19.

Hommel, B. (1998). Event files: Evidence for automatic integration of stimulus-response episodes. *Visual Cognition, 5*, 183-216.

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences, 8*, 494-500.

Hommel, B. & Colzato, L. (2004). Visual attention and the temporal dynamics of feature integration. *Visual Cognition, 11*, 483-521.

Huang, L. (2010). Visual working memory is better characterized as a distributed resource rather than discrete slots. *Journal of Vision, 10*, 8.

Irwin, D. E., & Yeomans, J. M. (1986). Sensory registration and informational persistence. *Journal of Experimental Psychology: Human Perception and Performance, 12*(3), 343.

Jones, G., Gobet, F., & Pine, J. M. (2007). Linking working memory and long-term memory: a computational model of the learning of new words. *Developmental Science, 10*, 853-873.

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica, 134*(3), 372-384.

Lavie, N. (2005). Load theory of selective attention and cognitive control. *Trends in Cognitive Science, 9*, 75-82.

Leboe, L. C., & Mondor, T. A. (2007). Item-specific congruency effects in nonverbal auditory Stroop. *Psychological Research, 71*(5), 568-575.

Loftus, G. R., & Irwin, D. E. (1998). On the relations among different measures of visible and informational persistence. *Cognitive Psychology, 35*(2), 135-199.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279-281.

Marks, L. E. (1987). On crossmodal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 384-394.

Marks, L. E., Ben-Artzi, E., & Lakatos, S. (2003). Crossmodal interactions in auditory and visual discrimination. *International Journal of Psychophysiology, 50*(1), 125-145.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits in capacity for processing information. *Psychological Review, 63*(2), 81-97.

Moeller, B., Frings, C., & Pfister, R. (2016). The structure of distractor-response bindings: Conditions for configural and elemental integration. *Journal of Experimental Psychology: Human Perception and Performance, 42*(4), 464-479.

Moeller, B., Pfister, R., Kunde, W., & Frings, C. (2016). A common mechanism behind distractor response and response-effect binding? *Attention, Perception, & Psychophsyics, 78*, 1074-1086.

Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton – Century – Crofts.

Olivers, C. N. L., Awh, E., & Van der Burg, E. (2016). The capacity to detect synchronous audiovisual events is severely limited: Evidence from mixture modeling. *Journal of Experimental Psychology: Human Perception and Performance, 42*(12), 2115-2124.

Parise, C. V., & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS ONE 4*(5), e5664.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia – a window into perception, thought and language. *Journal of Consciousness Studies, 8*(12), 3-34.

Rusconi, E., Kwan, B., Giordano, B. L., Umilta, C., & Butterworth, B. (2006). Spatial representation of pitch height: the SMARC effect. *Cognition, 99*(2), 113-129.

Sargent, J., Dopkins, S., Philbeck, J., & Chichka, D. (2010). Chunking in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 576.

Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews, 7*(3), 269-284.

Soto-Faraco, S. & Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance, 35*(2), 580-587.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*, 971-995.

Spence, C., & Deroy, O. (2012). Crossmodal correspondences: Innate or learned? *i-Perception, 3*(5), 316-318.

Spence, C., & Squire, S. (2003). Multisensory integration: maintaining the perception of synchrony. *Current Biology, 13*(13), R519-R521.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological monographs: General and applied, 74*(11), 1.

Stone, J. V., Hunkin, N. M., Porrill, J., Wood, R., Keeler, V., Beanland, M., Port, M., & Porter, N. R. (2001). When is now? Perception of simultaneity. *Proceedings of the Royal Society of London B: Biological Sciences, 268*(1462), 31-38.

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428*, 751-754.

Van der Burg, E., Awh, E., & Olivers, C. N. L. (2013). The capacity of audiovisual integration is limited to one item. *Psychological Science, 24*(3), 345-351.

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1053-1065.

Van Meeuwen, L. W., Jarodzka, H., Brand-Gruwel, S., Kirschner, P. A., de Bock, J. J., & van Merriënboer, J. J. (2014). Identification of effective visual problem solving strategies in a complex visual domain. *Learning and Instruction, 32*, 10-21.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598-607.

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature, 428*(6984), 748-751.

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature, 438*, 500-503.

Walker, P. (2012). Cross-sensory correspondences and cross talk between dimensions of connotative meaning: Visual angularity is hard, highpitched, and bright. *Attention, Perception, & Psychophysics, 74*, 1792–1809.

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space, and quantity. *Trends in Cognitive Sciences, 7*(11), 483-488.

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans. *Learning and Motivation, 14*, 406-432.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*(3), 638-667.

Wilbiks, J. M. P., & Dyson, B. J. (2013a). Effects of temporal asynchrony and stimulus magnitude on competitive audio-visual binding. *Attention, Perception, & Psychophysics, 75*(8), 1883-1891.

Wilbiks, J. M. P. & Dyson, B. J. (2013b). The influence of previous environmental history on audiovisual binding occurs during visual-weighted but not auditory-weighted environments. *Multisensory Research, 26*, 561-568.

Wilbiks, J. M. P., & Dyson, B. J. (2016). The dynamics and neural correlates of audiovisual integration capacity as determined by temporal unpredictability, proactive interference, and SOA. *PLoS ONE 11*(12), e0168304.

Wilbiks, J. M. P., Rioux, D. M., & Dyson, B. J. (in preparation). Facilitation of audiovisual integration capacity by illusory contours and decreasing presentation speed. *Manuscript in preparation*.

Zampini, M., Shore, D. I., & Spence, C. (2003). Audiovisual temporal order judgments. *Experimental Brain Research, 152*(2), 198-210.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*, 233-235.

Author's Note

B.J.D. was supported by an Early Researcher Award granted by Ontario Ministry of Research and Innovation. J. M. P. W. was supported by an Ontario Graduate Studentship and by a McCain Postdoctoral Fellowship. We thank Dominic M. Rioux for assistance with data collection in Experiment 4. Correspondence should be addressed to: Jonathan Wilbiks, Department of Psychology, Mount Allison University, 62 York Street, Sackville, New Brunswick, Canada, E4L 1E2. E-mail: jwilbiks@mta.ca